The Institute for Strategic Dialogue's Recommendations and Comments on Oversight Board Case 2023-023-FB-UA

Thank you to the Oversight Board for the opportunity to comment on case 2023-023-FB-UA**,** regarding a post in Polish targeting transgender people. Transgender rights and lives are currently under attack not just in Poland or the United States but around the world**.** Now more than ever, it is critical to support and implement policies and laws to protect transgender and gender-diverse people in every of facet society – including on social media platforms.

The Institute for Strategic Dialogue (ISD) is an independent, non-profit organization dedicated to safeguarding human rights and reversing the rising tide of polarization, extremism, and disinformation worldwide. Our work includes in-depth research and analysis identifying and tracking online manipulation, mis- and disinformation, hate, and extremism in real time. We also formulate, advocate and deliver evidence-based policy approaches and programming.

Transgender and gender-diverse people around the world face an increased risk of harm due to the many forms of discrimination they face daily, both online and offline. The situation in Poland is particularly dire – it is currently ranked 42nd out of 49 countries in the IGLA-Europe's 2023 Rainbow Europe Map[1] – but not unique. The year 2022 saw 327 reported murders of transgender and gender-diverse people across the world. There is also higher suicide risk for transgender people than cisgender people: a recent study showed that transgender people in Denmark had 7.7 times the rate of suicide attempts and 3.5 times the rate of suicide deaths compared to the rest of the population. Other studies by organizations such as The Trevor Project found that the share of LGBTQI+ youth who reported "seriously considering suicide" increased from 2020 to 2022.

These concerning figures and rise in transphobic legislation or movements in countries like Poland[2] and the United States have led several human rights advocacy groups to take action.[3] Social media platforms, which many use for healthy debate, information-gathering, and learning about new topics, have the duty to protect LGBTQI+ people on their platforms by not only developing comprehensive policies but also enforcing them correctly and uniformly. This is especially critical when widespread online hate speech can cause serious offline consequences, such as encouraging individuals to cause physical harm to

[1] The IGLA-Europe's Rainbow Europe Map reviews legal protections for LGBTQI+ people in European countries.
[2] Page 4-5 of the Polish Commissioner for Human Rights' 2019 report on the legal situation of non-heterosexual and transgender people in Poland highlights the growing problems in the country.
[3] According to the OHCHR, States have the legal responsibility to protect LGBTQI+ people from violence, cruel, inhuman, and degrading treatment, discrimination based on gender identity, and more – no matter who is perpetrating it (i.e., private individuals and companies).

themselves or others. In this case, the post could be considered a violation of Meta's Hate Speech Policy[4], Suicide and Self-Injury Policy[5], and Bullying and Harassment Policy[6].

Our submission seeks to address the Oversight Board's request for comments on Meta's policies and enforcement practices regarding hateful content targeting transgender people:

1. ***Speech, whether in spoken, written or visual form, that may be described by users as "humorous" or "satirical," but which may spread hate speech or other forms of inflammatory rhetoric.***

Previous ISD research has shown that internet memes or content that might be described as "humorous" or "satirical" by some users can be used strategically to obfuscate extremist narratives and convey hateful meanings through association rather than explicit argument. Additionally, extreme right-wing movements regularly use memes to condense radical ideologies into a more 'palatable' format that is easier to spread online and recruit and radicalize others. It also gives users an easy way to deflect any accusations of violating platform policies, spreading hateful or extremist ideologies, or targeting other users or groups online: users can claim the content was a "joke" or "satire." However, hateful content that uses humor, whether spoken, visual, or written, is still hateful content.

Recommendations:

- Meta should clarify in its Hate Speech Policy how reviewers determine whether content was intended to be satirical or not, and what that process looks like, including by providing indicative examples.
- Meta should invest in content moderation systems – whether human or through machine learning – that can catch the spread of extremist and hateful ideology through memes and "humorous" content (especially if the user posting it has strikes for other policy violating content or actions in the past) and curb inflammatory rhetoric.
- Meta should proactively invest in content moderation systems that operate in local languages, which are more adept at capturing the levels of nuance required to better identify content that propagates extremist and hateful ideology.

2. ***The risks associated with widespread hate speech targeting LGBTQI+ people on social media and Meta's human rights responsibilities in this context.***

Widespread hate speech targeting LGBTQI+ people online have serious offline consequences. In June 2023, ISD published a series of reports highlighting how anti-drag mobilization efforts (which frequently

[4] From Meta's Hate Speech Policy, Tier 2: "Targeting a person or group of people on the basis of their protected characteristic(s) with: Generalizations that state inferiority (in written or visual form) [including] mental health" and "expressions that a protected characteristic shouldn't exist."
[5] From Meta's Suicide and Self-Injury Policy: "We also remove content that identifies and negatively targets victims or survivors of suicide or self-injury seriously, humorously or rhetorically, as well as real time depictions of suicide or self-injury" and "Do not post: […] content that mocks victims or survivors of suicide."
[6] From Meta's Bullying and Harassment Policy: "Everyone is protected from […] calls for self-injury or suicide of a specific person, or a group of individuals.

amplify anti-trans talking points) are organized online and carried out offline. Our US report showed how online toxicity and hateful rhetoric can lead to offline aggression and verbal or physical assault. Our UK report showed how UK-based anti-drag activists were influenced by US activists and content online. In Poland, LGBTQI+ pride parades have ended in verbal or physical assault, with government officials parroting some of the rhetoric attacking LGBTQI+ people that is popular online, such as the "groomer" slur. Meta owes its LGBTQI+ users safe platforms where they can exercise their freedom of self-expression without fear of retaliation or hate speech.

Recommendations:

- Meta's policy teams need to be responsive to these kinds of trends and adapt policies and enforcement accordingly.
- Meta should also regularly brief moderators on emerging or spiking forms of hate and potential content violations.

3. ***Statements that encourage or applaud death by suicide as a form of hate speech, and whether Meta's policies and enforcement practices are sufficiently adequate to address them.***

Currently, Meta's Hate Speech Policy does not sufficiently address statements encouraging or applauding death by suicide of people with certain protected characteristics. The closest the policy gets to doing so is by prohibiting "expressions that a protected characteristic shouldn't exist." Similarly, the Suicide and Self-Injury Policy does not once refer to the Hate Speech Policy or address protected characteristics. While case 2023-023-FB-UA could technically fall under Hate Speech or Suicide and Self-Injury, the most relevant policy that was overlooked by Meta and reviewers was the Bullying and Harassment Policy, which states that "everyone is protected from […] calls for self-injury or suicide of a specific person, or a group of individuals."

Recommendations:

- Meta should bridge the gap between its Suicide and Self-Injury Policy and Hate Speech Policy by adding a clause in its Hate Speech Policy prohibiting posts alluding to, suggesting, or even outright stating that people with a protected characteristic(s) should die by suicide.
- Meta should give users the option to report a post for multiple violations. In this case, it might have been hard for a user to decide between which policy to prioritize using the existing user reporter tools: Hate Speech, Bullying and Harassment, or Suicide and Self-Injury. It would have also allowed Meta's reviewers to understand that this case was at an intersection of, and likely violating, multiple Meta policies.

4. ***Meta's policies and practices for reviewing multiple user reports involving the same piece of content.***

In the past, ISD has documented how Meta's "delays or mistakes in policy enforcement" have allowed for hateful and harmful content to spread through paid targeted ads. In the past couple of years, researchers and organizations have noted that these repeated delays or mistakes have extended beyond just ads, and that Meta's practices for reviewing violative content are not always entirely accurate. While 100% accuracy is unrealistic, in this case, there seemed to yet again be an inconsistency in the flagging

of the content to human reviewers and the amount of information sent to human reviewers to help inform their decision.

Recommendations:

- Meta should set a pre-determined number or percent of reports (no matter what policy they fall under) over a certain number of impressions or views, when reached, the content is automatically sent to a human reviewer. With this set policy in place, for example, if an Instagram post were to be reported 10 times (2 times for Harassment and Bullying, 5 times for Hate Speech, 3 times for Suicide and Self-Injury) for 100 views or impressions, Meta would automatically send it to a human reviewer – regardless of the virality or severity.
- Meta should invest in more human expertise to continue to finetune the balance between human moderation and automated moderation in its cross-check program.
- Meta should be transparent about how they uniformly and unbiasedly determine "high-impact content" in their cross-check system and how they choose what gets sent to human reviewers. Meta should provide a breakdown every quarter of the themes of cases that were sent to human reviewers.
- Meta should inform human reviewers that receive cases that have been reported under multiple policies which policies were selected by the users reporting. It is unclear whether the human reviewers knew that users also reported the post in case 2023-023-FB-UA for Hate Speech or just Suicide and Self-Injury.

5. ***Meta's account-level enforcement practices for users who repeatedly engage in anti-trans hate speech and harassment.***

Meta should have zero tolerance for users who repeatedly engage in anti-trans hate speech and harassment, especially if the user has been banned or suspended from Meta platforms before. Meta should ban IP addresses, phone numbers, or emails of repeat offenders to dissuade them from rejoining the platform.